Design and Implementation of New Learning Methods for Malicious URLs Phishing Detection

Synopsis Submitted to the

Swami Ramanand Teerth Marathawada University, Nanded

for the award of the degree of

Doctor of Philosophy

in the subject of

Computer Science and Engineering
Under the Faculty of Science and Technology.

Submitted by

Rachana S. Potpelwar

under the guidance of

U. V. Kulkarni
Professor, Computer Science and Engineering

Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded (M.S.) India.

2022

Contents

1 Introduction **2**
1.1 Importance 2
1.2 Motivation 3
1.2.1 Challenges with the Existing System 3
1.3 Objectives 4

2 Literature Survey **4**

3 Methodology **10**
3.1 Feature Characterization and Preprocessing 10
3.1.1 URL Based Features 10
3.1.2 HTML-related features 11
3.2 Model Evaluation 11
3.2.1 Soft Computing Tools 11
3.2.2 Classification Methods 11
3.2.3 Ensemble Learning 13
3.2.4 Deep Learning and TensorFLow 13
3.3 Data Collection Procedure 14
3.4 Softwares 14

4 Importanceof Study/Social Application **14**
4.1 National and International status 14
4.2 Significance of the study 15

5 Proposed work plan/Formulation and structure of study **15**

Abstract of proposed work plan/problem: The web users who use online services, e-banking and online shopping is increasing due to the comfort flexibility. So, there is growth of the cyber attackers, phishers, and hackers to gain the confidential and financial information of the web users. There are many phishing datasets are available which has both phishing URLs and legitimate URLs. In this domain there is active research, applying advance learning techniques on the datasets and the features can help the online services to reduce the threat of the intruders.
The dataset have different features and there is need to identify with the help of features-based techniques. There are many data prepossessing steps to reduce and improve data. Lexical based method is one approach to process the data and apply advance learning methods for model evaluation.[**1**]
However there are many phishing detection techniques, for detecting the attackers. Here, the work is to design and optimize these learning methods and techniques to overcome with the challenges of the security.

1 Introduction

As per the rapid increase of web users who uses online services, e-banking and online shopping is increasing due to the comfort flexibility and ease of use. So, the cyber attackers, phishers, and hackers to gain the confidential and financial information of the web users has increased. There are many different types of the phishing attacks to name few, Emails phishing, HTTPs phishing, spear phishing, smishing, angular phishing, pop-up phishing, clone- phishing, Evil twin, watering hole phishing, whaling, Business Email Commerce (BEC), vishing, smishing, snowshoeing and many. So there is the need to identify the phishing websites and legitimates sites in order to prevent ourselves from losing the confidential data or any banking transaction theft.

1.1 Importance

The increasing popularity of internet-based systems, the security challenges are the primary cause of cyber fraud and identity theft. So, it is important for checking the

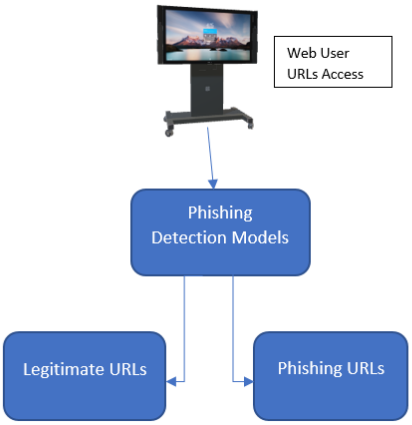

Figure 1: Phishing URLs Detection

legitimacy of visited web pages and is a crucial task to secure costumers' identities and prevent phishing attacks.

1.2 Motivation

Phishing has been recognized as the easiest and widespread cyber criminality threat and hackers do not crack any complex cypher code neither breach hard firewalls just, they send emotional messages and urging their personal credentials [**5**]. Many phishing websites are launching every 20 seconds. So, all this require the detection and prevention of such cyber criminality attacks. Since it is very important to detect phishing attacks. Main motivation is to reduce the cybercrime.

1.2.1 Challenges with the Existing System

Everyday millions of new websites are being developed that have login portals to extract credentials from users, as these websites are in large volume, it is becoming challenging to verify their credibility. According to a report published by Dofo, more than 5.16 million domain names were registered in the month of April 2020 [**7**]. As the number of users is increasing, it becomes easy for attackers to lure more users. So, it is every second challenge to overcome with this existing systems.

1.3 Objectives

The main objective of the research is to identify high accuracy feature based phishing URLs and reduce the cyber threat. With the help of advance learning algorithms, optimize the accuracy of identifying the phishing URLs.

Web based phishing URLs detection techniques, in these techniques the user classify the URLs for phishing and legitimate using optimized learning methods.[**8**]

2 Literature Survey

Review of literature and development in the subject (Previous work done in the relevant area)

- Brij B. Gupta a,b, Krishna Yadav a, Imran Razzak c, Konstantinos Psannis d, Arcangelo Castiglione e, Xiaojun Chang "A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment."[**2**] The work has proposed the hidden knowledge from the proposed SVM model by adopting a related method embedded, the extracted rules into a browser extension named Phish Detector. To make the proposed method more functional and easier to use. Evaluating of the implemented browser extension indicates that it can detect phishing attacks in internet banking with high accuracy and reliability. Phish Detector can detect zero-day phishing attacks too. The proposed model can detect phishing pages in internet banking with accuracy of 99.14 % true positive and only 0.86 % false negative alarm. The novelty of the paper is anti-phishing approach that can detect phishing attacks without utilizing many resources and bandwidth. The proposed method uses Dom features to improve results. The limitations are feature set are entirely depending on the web page content, image cannot be determined with this proposed method.
- Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). "Phishing detection based associative classification data mining." The research finding is phishing lifestyles, non-technical approach to handle phishing, technical approaches to handle phishing. The associate classification extracts

classification contains sample if-then rules with a high degree of predict accuracy. The novelty MCAC was able to produce much label rules from the phishing data generates rules assoc. With new class called suspicious. The strengths are ensemble methods for study model to get 97.3 accuracy rate CNN is used to improve the visual feature. [**3**]

- Peng Yang, Guangzhen Zhao, and Peng Zeng, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning," the multidimensional feature phishing detection (MFPD) approach based on a fast detection method by using deep learning is proposed. In the first step, character sequence features of the given URL are extracted and used for quick classification by deep learning. Specially, the CNN (convolutional neural network) is used to extract local correlation features through a convolutional layer. In a URL, each character may be related to nearby characters.

  A phishing website is likely to mimic the URL of a legitimate website by changing or adding some characters. This can cause the sequential dependency of the phishing URL to be different from the phishing URL. The LSTM network can effectively learn the sequential dependency from character sequences. Therefore, the LSTM (long short-term memory) network is employed to capture context semantic and dependency features of URL character sequences, and at finally SoftMax is used to classify the extracted features. The authors have concluded that the MFPD approach is effective with high accuracy, low false positive rate and high detection speed.[**19**]

- Kang Leng Chiew, Choon Lin Tan, KokSheik Wong, Kelvin S.C. Yong, Wei King Tiong "A new hybrid ensemble feature selection framework for machine learning- based phishing detection system." The author has collected the datasets from different sources i.e., phish tank, open phish, Alexa, common crawl, GNU Wget, pythons script, UCI .The different research findings involved in this paper are filter measures rankings, performance of basic features, runtime analysis comparison. HEFS performs best with RF class where baseline features correctly distinguish 94.6 % of phishing. HEFS is proposed to find a subset of features for

utilization in ml-based phishing detection with HEFS a novel algo called CDFG is proposed to automatically determine the optimal number of features. [**9**]

- Yan Ding, Nurbot Lekstshan, Kepin Li, Whshour Slamu "A keyword-based combination appraoch for detecting phishing webpages."

  The following datasets are used Phish tank, Yahoo, URLB, Dmoz. The different classification methods used are SVM, Bayes, LR (L1, L2) search engine based detection method Heuristic rule based detection method, LR classifiers based detection phase, SHLR data method Being search engine Beyondb, PhishDef, Bigdata, SHLR methods has used to find the phishing websites from the webpages. Obfuscation techniques can filter legal webpages and detect phishing webpages by adopting escape technology Real time detection can be satisfied with the SHLR method which reads the false positive rate caused by the lack of rules and improves the adaptability and accuracy of the method the browser plugins have to be developed based on the SHLR. This paper combines the advantage of the three detection methods and proposes the SHLR phishing webpage data method. Rule based detection method does not need to extract features and perform efficient matches of string consumes time to extract features lower detection efficiency which is not conducive to real time detection.[**25**]

- Moghimi, Mahmood, and Ali Yazdian Varjani. "New rule-based phishing detection method." The datasets used are legitimate web pages from yahoo directory service and internet banking websites. Two features' sets have given i.e., improve the performance, prevent data loss in internet banking. The results are Number of rules acuracy-98.58 error- 1.41, Sensitivity 1, F1-measure 0.98, rules extracted directly from training dataset accuracy 97.89, error- 2.17, sensitivity 0.9, f1 measure - 0.98.[**4**]

- Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). "Machine learning based phishing detection from URLs." The research finding NLP based features have better performance than word vectors with the average rate of 10.86 for hybrid 2.24 and hybrid vs word vec 13.14. The conclusion NLP based feature have better performance than word

vectors with average rate of 10.86 percent. The novelty of the work is to increase the effectiveness, the author has compared the performance different between NLP, and hybrid features.[**6**]

- Li, Yukun, et al. "A stacking model using URL and HTML features for phishing webpage detection." The datasets used are 50K-PD contains 49974 webpages,50-IPD contains 53103 webpages. Stacking model is derive by coming GBDT XGBoost and lightGBM to detect phishing webpages. Since enable methods are used to calculate diver and effective stacking model. Language independence, huge sized of phishing and legitimate data, real time execution, detection of new websites, independence from third party service use of feature rich classifiers [**7**].
- Hannousse, A., & Yahiouche, S. (2021). "Towards benchmark datasets for machine learning based website Phishing detection: An experimental study." Filter based ranking with incremental removal of less important features the performance up to 96.83% better than wrapper methods. The features used are URL based features, content based features, external based features[**10**].
- Ali, W., & Malebary, S. (2020). "Particle swarm optimization-based feature weighting for improving intelligent phishing website detection." The work is based on feature weight selection. In this paper, training phase has a set of phishing and legitimate websites is used to train and build an intelligent detection model. The Particle swarm optimization (PSO) algorithm is used in this paper since it achieves fast convergence in numerous optimization problems. After feature weighting based on PSO, six popular machine learning algorithms were used i.e., back - propagation neural network (BPNN), support vector machine (SVM), k-nearest neighbor (KNN), decision tree (C 4.5), random forest (RF) and naive Bayes (NB). The classification accuracy was outstanding after applying PSO- based feature weighting method. [**13**]
- Rahman, S. S. M. M., Islam, T., & Jabiullah, M. I. (2020). "PhishStack: evaluation of stacked generalization in phishing URLs detection." The features used are html JavaScript based features, abnormal based features, domain-based features, address bar-based features.

Maximum 95.61% of accuracy attained from random forest individually with minimum error rate, it has been noticed that random forest and MLP provide better precision and recall but stacked generalization performs better in terms of overall accuracy. [**14**]

- Lin, G., Sun, N., Nepal, S., Zhang, J., Xiang, Y., & Hassan, H. (2017). "Statistical twitter spam detection demystified: performance, stability and scalability." The data used for this paper is the trend micro's web reputes service 65 million spam tweets. R programming language (version 3.2.3) with caret package and do Parallel Package are used. RF and c5-0 are ideal candidate for real world spam detection based on parallel computing techniques are applied. ML algorithms are used to study classifiers performance stability based on tweets datasets. The novelty of the work, with different number of CPU cores, observed NN parallel computational resources impact algorithms suitable for a Cartan problem size. [**15**]
- Fang, Y., Zhang, C., Huang, C., Liu, L., & Yang, Y. (2019). "Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism." The manual extraction of features is avoided because we adopted the deep learning method. The THEMIS model obtains a promising result. [**16**]
- Zhu, E., Chen, Y., Ye, C., Li, X., & Liu, F. (2019). "OFS-NN: an effective phishing websites detection model based on optimal feature selection and neural network." The new FVV index was firstly defined to evaluate the impact of sensitive features on phishing detection. The dataset used are from UCI Phishing website dataset. [**17**]
- Yuan, J., Chen, G., Tian, S., & Pei, X. (2021). "Malicious URL detection based on a parallel neural joint model." In this work, the datasets are collected from phishing webpages and benign webpages from Phish Tank and Alexa websites respectively. To better learn the information from chars, embedded vector seq., Ind RNN is selected. And IN capsNET pooling layer is removed from CNN to fully utilized each feature. The novelty of the work is traditional ML algorithm need to manually extract the features of URLs, by using RNN algorithm more appropriate

URL classification is achieved. [**18**]

- Yang, P., Zhao, G., & Zeng, P. (2019). "Phishing website detection based on multidimensional features driven by deep learning." Real life dataset collected from internet DATA1 and DATA2. The features used are URL statistical features, webpage code features, webpages text features. Multidimensional feature phishing detection approach based on fast detection method by using deep learning with less amount of time. Optimization strategy DCDA (Dynamic category Decision algorithm) is used for balancing accuracy and speed. Also, fast detection method is used (through DL) which does not require third party assistant or prior knowledge. The limitation of the work, as multidimensional features are used for detection, the size of dataset is very large and may take more time to detect.[**19**]

- Feng, J., Zou, L., Ye, O., & Han, J. (2020). "Web2Vec: Phishing Webpage Detection Method Based on Multidimensional Features Driven by Deep Learning." The datasets used phishing webpages and benign webpages from Phish Tank and Alexa websites respectively. The features used are @ symbol in URL, IP based URL, URI with hexadecimal code, long URL length, multiple slash in the URL path. The model uses the representation learning technology in NLP to comprehensively learn the representation of webpages from the URL, page content, and DOM structure; then construct a multi-channel hybrid deep learning network to extract the deep hidden features of the webpages and then use the attention mechanism to strengthen the influence of important features; finally, the feature extraction results of different channels are fused for classification prediction. Four sets of experiments verified the classification results of the Web2Vec model from different angles. Four sets of experiments verified the classification results of the Web2Vec model from different angles. The accuracy can be improved with DOM structure. [**20**]

- Adebowale, M. A., Lwin, K. T., Sanchez, E., & Hossain, M. A. (2019). "Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text." ANFIS based robust

scheme using the integrated features of the text, images and frames for web-phishing detection and protection. Classification on text features accuracy 98.55, frame feature 98.06, image feature 97.20, hybrid feature-98.30. [**21**]

- Vrbančič, G., Fister Jr, I., & Podgorelec, V. (2020). "Datasets for phishing websites detection." The datasets used are phish tank and Alexa. Two CSV files have been generated with python script to make the files easy. Applied on 2-fold, 5-fold and 10-fold, integration of hybrid features that have been extracted from text, images and frames and that are then used to develop a robust ANFIS solution. [**22**]

3 Methodology

3.1 Feature Characterization and Preprocessing

Feature selection and filtering Phishing Websites Features. The features in the dataset play crucial role for predicting phishing websites. There are different types of the features.[**28**]
1. URL Based feature
2. Page Based features
3. HTML and JavaScript Based Features
4. Abnormal Based Features

3.1.1 URL Based Features

IP Address : Usually, the domain names of phishing websites have IP address. A binary value is used to indicate whether the domain name of a URL is an IP address or not.[**23**]
https : https is a secure "http" data transfer method that provides authentication and encrypted communication.
URL length information : Length information includes the number of the characters in a URL and its domain name respectively. [**29**]
Number of dots in domain name : The number of "." in domain name is used

to represent this feature.
Similar target brands: Phishing webpage designers usually limitate the URLs of target pages to confuse users. For example, quite a few phishing ones modify "paypal" as "paypail" to confuse users.

3.1.2 HTML-related features

Number of internal and external links : An internal or external link refers to the one has the same or differnt domain name as the URL.
Empty link: There are two types of empty links. The number of empty links in the webpage is the feature
Internal and External resources : The number of internal and external resources has to be calculated as features. The external resources in phishing webpages are link, img, script and nonscript
Length of HTML content: To cheat the users the length of the HTML is less in phishing webpages than in legitimate webpages.
Alarm Window: Alarm window is a pop up window generated by phishing websites for sensitive information. [**26**]

3.2 Model Evaluation

3.2.1 Soft Computing Tools

In recent times, engineers have very well accepted soft computing tools such as Fuzzy Computing, Neuro-Computing, Evolutionary Computing, Probabilistic Computing, and Immunological Computing etc., for carrying out various numerical simulation studies. In last two decades, these tools independently as well as in hybrid forms has been successfully applied to varieties of problems.

3.2.2 Classification Methods

1. **Multiclass Classification**
Multiclass classification involves more than two classes. There are two approaches. A simple approach is one-versus-all (OVA). All-versus-all (AVA) is an alternative approach that learns a classifier for each pair of classes.

Error-correcting codes can be used to improve the accuracy of multiclass classification, not just in the previous situations, but for classification in general. Error-correcting codes were originally designed to correct errors during data transmission for communication tasks. For such tasks, the codes are used to add redundancy to the data being transmitted so that, even if some errors occur due to noise in the channel, the data can be correctly received at the other end. For multiclass classification, even if some of the individual binary classifiers make a prediction error for a given unknown tuple, we may still be able to correctly label the tuple.[**24**]

2. **Semi-Supervised Classification**

Self-training is the simplest form of semi-supervised classification. It first builds a classifier using the labeled data. The classifier then tries to label the unlabeled data. The tuple with the most confident label prediction is added to the set of labeled data, and the process repeats.

Cotraining is another form of semi-supervised classification, where two or more classifiers teach each other. Each learner uses a different and ideally independent set of features for each tuple. Consider web page data, for example, where attributes relating to the images on the page may be used as one set of features, while attributes relating to the corresponding text constitute another set of features for the same data.

3. **Active Learning**

Active learning is an iterative type of supervised learning that is suitable for situations where data are abundant, yet the class labels are scarce or expensive to obtain. The learning algorithm is active in that it can purposefully query a user (e.g., a human oracle) for labels. The number of tuples used to learn a concept this way is often much smaller than the number required in typical supervised learning.

4. **Transfer Learning**

The classification task is to automatically label the reviews as either positive or negative. This task is known as sentiment classification. We could examine each review and annotate it by adding a positive or negative class label. The labeled reviews can then be used to train and test a classifier to label future reviews of the product as either positive or negative. The manual effort involved in annotating the review data can be expensive and time consuming. Transfer learning aims to extract the knowledge from one or more source tasks and apply the knowledge to

a target task.

3.2.3 Ensemble Learning

The ensemble methods are classified into two main categories. Bagging and Boosting are two main categories of ensemble methods. In bagging technique, which is the example for Random Forest, the process of training is done on the random selection of splits and prediction that are based on majority vote. Whereas the boosting main focus is on the misclassified samples. Ad boost and Gradient tree boosting are the examples of the boosting techniques [**36**, **37**]

1. **Random Forest Classification**

A random forest is a set of decision trees built on random samples with a different policy for splitting a node.

2. **Gradient tree boosting :**

The main aim of the gradient tree boosting is to minimize the loss function. It executes step by step in order to achieve more accuracy.

3. **AdaBoost Classifier :**

Adaptive Boosting algorithm is the ensemble method which is the collection of the different classifier and taking the average vote of the classifier to execute good and improved results. Ada boost algorithm focus is on analysis of misclassified samples.

4. **Voting classifier :** It is not a classifier but a wrapper for a set of different ones that are trained and evaluated in parallel. Voting Classifier supports two types of voting.

Hard Voting : In hard voting, the predicted output class is a class with most votes i.e., the class which had the highest probability of being predicted by each of the classifiers.

Soft Voting : In soft voting, the output class is the prediction based on the average of probability given to that class.

3.2.4 Deep Learning and TensorFLow

Deep learning has become very famous in the last few decades, many applications that are changing the way with many electronic (and non-electronic) systems. Speech, text, and image recognition, autonomous vehicles and intelligent bots are

common applications normally based on deep learning models and have outperformed any previous classical approach[**38**]. An artificial neural network (ANN) or simply neural network is a directed structure that connects an input layer with an output one [**30-34**]
TensorFlow is a computational framework created by Google and has become one of the most diffused deep-learning toolkits. It can work both CPUs and GPUs and already implements most of the operations and structures required to build and train a complex model.[**38**]

3.3 Data Collection Procedure

The dataset is collected from UCI repository, Phishing website dataset, the dataset contains 31 columns with 30 features and 1 target. The dataset has 2456 observations. There are many other sources of phishing dataset like PhishTank Dataset, Alexa[**22**].

3.4 Softwares

Google Colab, Python Jupiter Notebook for analysis. ITK toolkit, VTK toolkit for visualization. For big data analysis Apache Spark 2.2.0.

4 Importanceof Study/Social Application

4.1 National and International status

Human factor influence how individual interact with cybersecurity systems and are the cause of many successful cyberattacks. Phishing has been recognized as the easiest and widespread cyber criminality threat, hackers do not crack any complex cipher code neither breach hard firewalls, just they send emotional messages and urge users personal credentials [**39**]. Different optimized learning methods are introduced to classify the phishing and legitimate URLs. Still, phishing detection URLs is a open challenge for developing intelligent security mechanism

4.2 Significance of the study

To ensure a high level of security in all the fields of phishing detection of URLs [**35**]. Malware is a challenging cybersecurity threat, as it is often bundled with legitimate software. The other approaches, the problem of automating fake news detection, email threat landscape and many. So, to reduce the threat of phishing, the study aims at designing the new trends and optimized techniques in advance learning methods [**40**].

5 Proposed work plan/Formulation and structure of study

- **Year-wise plan of work and targets to be achieved Year Plan.**

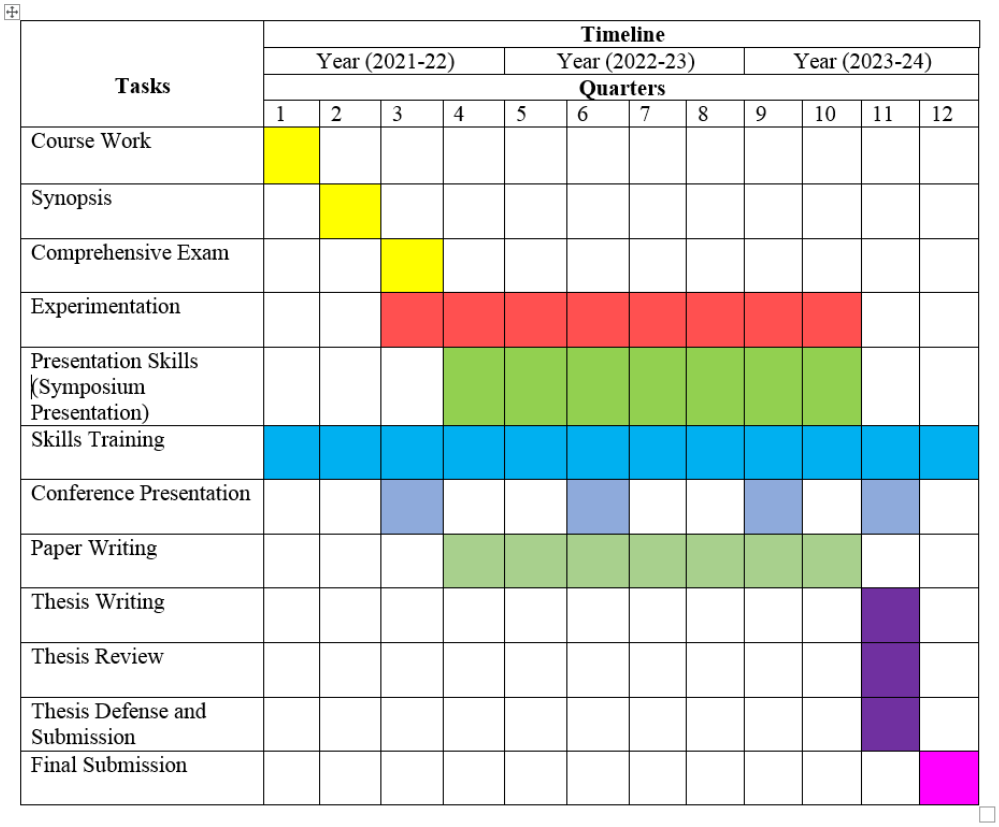

Figure 2: Timeline of Proposed Research Work

References

[1] Khonji, Mahmoud, Youssef Iraqi, and Andrew Jones. "Phishing detection: a literature survey." IEEE Communications Surveys and Tutorials 15.4 (2013): 2091-2121.

[2] Gupta, B. B., Yadav, K., Razzak, I., Psannis, K., Castiglione, A., & Chang, X. (2021). "A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment." Computer Communications, 175, 47-57.

[3] Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014)."Phishing detection based associative classification data mining." Expert Systems with Applications, 41(13), 5948-5959.

[4] Moghimi, Mahmood, and Ali Yazdian Varjani. "New rule-based phishing detection method." Expert systems with applications 53 (2016): 231-242.

[5] Tan, C. L., Chiew, K. L., Yong, K. S., Abdullah, J., & Sebastian, Y. (2020). "A graph-theoretic approach for the detection of phishing webpages." Computers & Security, 95, 101793.

[6] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). "Machine learning based phishing detection from URLs." Expert Systems with Applications, 117, 345-357.

[7] Li, Yukun, et al. "A stacking model using URL and HTML features for phishing webpage detection." Future Generation Computer Systems 94 (2019): 27-39.

[8] Orunsolu, A. A., Sodiya, A. S., & Akinwale, A. T. (2019). "A predictive model for phishing detection." Journal of King Saud University-Computer and Information Sciences.

[9] Chiew, K. L., Tan, C. L., Wong, K., Yong, K. S., & Tiong, W. K. (2019). "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system." Information Sciences, 484, 153-166.

[10] Hannousse, A., & Yahiouche, S. (2021). "Towards benchmark datasets for machine learning based website phishing detection: An experimental study." Engineering Applications of Artificial Intelligence, 104, 104347.

[11] Alsariera, Y. A., Adeyemo, V. E., Balogun, A. O., & Alazzawi, A. K. (2020). "AI meta-learners and extra-trees algorithm for the detection of phishing websites." IEEE Access, 8, 142532-142542.

[12] V. N. Uzel, E. S. Essiz, and S. A. Ozel, "Using fuzzy sets for detecting cyber terrorism and extremism in the text," in Proc. Innov. Intell. Syst. Appl. Conf. (ASYU), Oct. 2018, pp. 1–4.

[13] Ali, W., & Malebary, S. (2020). "Particle swarm optimization-based feature weighting for improving intelligent phishing website detection." IEEE Access, 8, 116766-116780.

[14] Rahman, S. S. M. M., Islam, T., & Jabiullah, M. I. (2020). "PhishStack: evaluation of stacked generalization in phishing URLs detection." Procedia Computer Science, 167, 2410-2418.

[15] Lin, G., Sun, N., Nepal, S., Zhang, J., Xiang, Y., & Hassan, H. (2017). "Statistical twitter spam detection demystified: performance, stability and scalability." IEEE access, 5, 11142-11154.

[16] Fang, Y., Zhang, C., Huang, C., Liu, L., & Yang, Y. (2019). "Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism." IEEE Access, 7, 56329-56340.

[17] Zhu, E., Chen, Y., Ye, C., Li, X., & Liu, F. (2019). "OFS-NN: an effective phishing websites detection model based on optimal feature selection and neural network." IEEE Access, 7, 73271-73284.

[18] Yuan, J., Chen, G., Tian, S., & Pei, X. (2021). "Malicious URL detection based on a parallel neural joint model." IEEE Access, 9, 9464-9472.

[19] Yang, P., Zhao, G., & Zeng, P. (2019). "Phishing website detection based on multidimensional features driven by deep learning." IEEE access, 7, 15196-15209.

[20] Feng, J., Zou, L., Ye, O., & Han, J. (2020). " Web2Vec: Phishing Webpage Detection Method Based on Multidimensional Features Driven by Deep Learning." IEEE Access, 8, 221214-221224.

[21] Adebowale, M. A., Lwin, K. T., Sanchez, E., & Hossain, M. A. (2019). "Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text." Expert Systems with Applications, 115, 300-313.

[22] Vrbančič, G., Fister Jr, I., & Podgorelec, V. (2020). "Datasets for phishing websites detection." Data in Brief, 33, 106438.

[23] Maria Sameen 1, Kyungyun HAN 2, and Seong oun hwang, "PhishHaven — An Efficient Real-Time AI Phishing URLs Detection System," Received April 10, 2020, accepted April 24, 2020, date of publication April 30, 2020, date of current version May 15, 2020.

[24] Xiao, X., Xiao, W., Zhang, D., Zhang, B., Hu, G., Li, Q., & Xia, S. (2021)."Phishing websites detection via CNN and multi-head self-attention on imbalanced datasets." Computers & Security, 108, 102372.

[25] Ding, Y., Luktarhan, N., Li, K., & Slamu, W. (2019). "A keyword-based combination approach for detecting phishing webpages." computers & security, 84, 256-275.

[26] Saravanan, P., & Subramanian, S. (2020). "A framework for detecting phishing websites using GA based feature selection and ARTMAP based website classification." Procedia Computer Science, 171, 1083-1092.

[27] Azeez, N. A., Misra, S., Margaret, I. A., & Fernandez-Sanz, L. (2021). "Adopting automated whitelist approach for detecting phishing attacks." Computers & Security, 108, 102328.

[28] Gualberto, E. S., De Sousa, R. T., Thiago, P. D. B., Da Costa, J. P. C., and Duque, C. G. (2020). "From feature engineering and topics models to enhanced prediction rates in phishing detection." Ieee Access, 8, 76368-76385.

[29] Cohen, D., Naim, O., Toch, E., and Ben-Gal, I. (2021). "Website categorization via design attribute learning." Computers and Security, 107, 102312.

[30] Yaseen, Qussai. “Spam email detection using deep learning techniques.” Procedia Computer Science 184 (2021): 853-858.

[31] Gallo, L., Maiello, A., Botta, A., and Ventre, G. (2021). “2 Years in the anti-phishing group of a large company.” Computers and Security, 105, 102259.

[32] Bai, T., Bian, H., Salahuddin, M. A., Abou Daya, A., Limam, N., and Boutaba, R. (2021).“Rdp-based lateral movement detection using machine learning.” Computer Communications, 165, 9-19.

[33] GitHub. “Implementation for the Usage of Google Safe Browsing APIs,”.Accessed:2019.[Online]. Available:https://github.com/google/safebrowsing

[34] V. N. Uzel, E. S. Essiz, and S. A. Ozel,“Using fuzzy sets for detecting cyber terrorism and extremism in the text,” in Proc. Innov. Intell. Syst. Appl. Conf. (ASYU), Oct. 2018, pp. 1–4.

[35] Huang, S. Hao, L. Invernizzi, Y. Fang, C. Kruegel, and G. Vigna, “Gossip: Automatically identifying malicious domains from mailing list discussions,” in Proc. ACM Asia Conf. Comput. Commun. Secur. (ASIA CCS), Abu Dhabi, United Arab Emirates, Apr. 2017, pp. 494-505.

[36] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri, “Machine learning based phishing detection from URLs Contents,” lists available at Science Direct Expert Systems With Applications journal homepage www.elsevier.com/locate/eswa

[37] F. Vanhoenshoven, G. Nápoles, R. Falcon, K. Vanhoof, and M. Köppen, “Detecting malicious URLs using machine learning techniques,” in Proc. IEEE Symp. Ser. Comput. Intell. (SSCI), Dec. 2016, pp. 1-8.

[38] J. Saxe, R. Harang, C. Wild, and H. Sanders, “A deep learning approach to fast, format-agnostic detection of malicious Web content,” in Proc. IEEE Symp. Secur. Privacy Workshops (SPW), San Francisco, CA , USA, Aug. 2018, pp. 8-14.

[39] M. Junger, L. Montoya, and F.-J. Overink, "Priming and warnings are not effective to prevent social engineering attacks," Comput. Hum. Behav.,vol. 66, pp. 75-87, Jan. 2017.

[40] Yar, M., & Steinmetz, K. F. (2019). Cybercrime and society. Sage.

Signature

Mrs. Rachana S. Potpelwar
Research Scholar

Signature

Dr. Uday V. Kulkarni
Research Supervisor